DOCUMENT RESUME

ED 336 408 TM 017 151

AUTHOR Doyle, Teresa F.; Lin, Thung-Rung
TITLE Criterion Validation of a Written Multiple-Choice Test of Spanish/English Bilingual Skills.
PUB DATE Jun 91
NOTE 19p.; Paper presented at the Annual Conference of the International Personnel Management Association Assessment Council (15th, June 26, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adults; Bilingualism; Bilingual Teachers; English; Holistic Evaluation; Job Performance; Language Tests; *Multiple Choice Tests; Objective Tests; Occupational Tests; Predictive Validity; *School Personnel; Scoring; Spanish; *Spanish Speaking; Test Format; *Test Validity; Translation; *Written Language
IDENTIFIERS Validation Verification and Testing Techniques

ABSTRACT

Supervisory performance appraisals may be of limited utility in the validation of bilingual tests because incumbents are often hired to be the only employee in a unit who possesses the skills necessary to do the job. In an effort to provide criterion-related validity for four equivalent forms of a Spanish/English bilingual test for school district employees, 63 volunteer employees in a large metropolitan school district, who received a pay differential for bilingual skills, were asked to translate, free-form, a passage into written Spanish from written English. This same passage was subsequently included as part of a multiple-choice test in which the examinees selected the correct translations of sentences or phrases from among four options presented to them. The free-form translations were scored by four subject-matter experts (SMEs) using a holistic scoring method developed by E. M. White (1986) and R. A. Honey (1990). SME scores were then averaged and correlated with the objective test scores for each volunteer. The relationship between objective test scores and holistic evaluations was used to establish a passing point for the objective tests. The strength of the relationship observed between the objective test scores and raters' evaluations supports the conclusion that the holistic scoring of translations can serve as an adequate validation method for the objective test. One figure and three tables present study data. An appendix contains the holistic scoring guide. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TERESA F. DOYLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED336408

CRITERION VALIDATION OF A WRITTEN MULTIPLE-CHOICE TEST OF SPANISH/ENGLISH BILINGUAL SKILLS

Teresa F. Doyle
Personnel Services
San Diego County Office of Education

Thung-Rung Lin
Personnel Selection Branch
Personnel Commission
Los Angeles Unified School District

Paper presented at the Fifteenth Annual Conference of the International Personnel Management Association Assessment Council, June 26, 1991.

TM017151

BEST COPY AVAILABLE

ABSTRACT

Supervisory performance appraisals may be of limited utility in the validation of bilingual tests because incumbents are often hired to be the only employee in a unit who possesses the skills necessary to do the job. In an effort to provide criterion-related validity for four equivalent forms of a Spanish/English bilingual test, volunteer employees receiving a pay differential for bilingual skills were asked to translate, free-form, a passage into written Spanish from written English. This same passage was subsequently included as part of a multiple choice test in which testees selected the correct translations of sentences or phrases from among four options presented to them. The free-form translations were scored by subject matter experts (SMEs) using a holistic scoring method (White, 1986; Honey, 1990). SME scores were then averaged and correlated with objective test scores for each volunteer, resulting in $r = .62$.

CRITERION VALIDATION OF A WRITTEN MULTIPLE-CHOICE TEST
OF SPANISH/ENGLISH BILINGUAL SKILLS

Introduction

Preliminary results from the recent census indicate that California is continuing on a rapid course to great ethnic diversity among its citizens (CCSCE, 1990). As a consequence, public agencies are increasingly presented with the task of providing services to individuals for whom English is a foreign language, as many of the newcomers to California are new to the United States as well (CCSCE, 1990). Included in this task are the needs to develop and validate testing procedures for selecting employees with bilingual proficiency, and questions regarding the tests' utility for identifying the adequately qualified are not simply answered.

The demonstration of criterion-related validity, of course, depends upon the selection of an appropriate criterion. Supervisory performance evaluations are commonly-used criteria for the purpose of test validation, but may have less than the desired utility in some instances because of the supervisor's inability to make an informed assessment of the incumbent's performance due to the highly specialized knowledge required to perform the duties under study. It has been observed, for example, that supervisors of technical and medical positions have difficulty in providing accurate appraisals of the abilities and performance of their professional subordinate staff (Cascio, 1986). This problem results in little or no differentiation among appraisees, and performance evaluations subject to this limitation would be poorly correlated with test scores.

Supervisory performance appraisals may be singularly unhelpful in the validation of bilingual tests because an incumbent is oftentimes hired to be the only employee in a unit who possesses the skills necessary to do the

job; monolingual supervisors are left to deduce, from a lack of complaints, that incumbents are performing satisfactorily. Meanwhile, those on the receiving end of the incumbent's translation who may wish to complain are unable to do so because they don't speak the supervisor's language. Furthermore, the supervisors may be totally unable to distinguish relative quality of performance, a critical element for demonstrating test validity when performance appraisals are used as the criterion.

For these reasons, it was decided to develop another criterion when four equivalent forms of a written test of Spanish/English bilingual proficiency were constructed for a large, metropolitan school district. The tests would be utilized as the initial tool in the selection of bilingual classified (that is, noncertificated) employees, including clerks, secretaries, and personnel analysts. Typically, incumbents of these positions utilize their biliteracy skills on the job to translate documents or correspondence from written English into Spanish and vice-versa; that is, little skill for original composition is usually required.

The Spanish/English bilingual proficiency written tests are multiple choice format. The debate regarding the use of multiple choice tests for the measurement of language writing ability has long raged (see Quellmalz, 1986), with opponents suggesting that writing samples must be obtained in order to adequately sample constructs related to writing. However, though a test format incorporating freehand translation tasks would have been more like the tasks incumbents are expected to perform on the job, the very large numbers of applicants tested on a continuing basis made such a format impractical because of the requirement for obtaining qualified raters to evaluate the translations. Even if qualified raters existed in great supply, the cost associated with administrations of such a test to large

numbers of people would seriously diminish its utility, as others have found in similar situations (Lawton and Showers, 1980). Moreover, translations and original writing differ in what may be a critical aspect insofar as test format is concerned, as translations do not require the conceptual creativity demanded by original composition which is so difficult to assess with multiple choice. On the other hand, it was extremely desirable to obtain empirical verification that a person's abilities as measured by the multiple choice test were significantly and positively correlated with his or her ability to make freehand translations.

It was decided to conduct a concurrent validity study in which current incumbents of bilingual classified positions took two alternate forms of the multiple choice test and also completed two freehand translations. The translations were then evaluated by subject matter experts (SMEs) who assigned them numerical scores, and these scores were correlated with the scores obtained by the individuals on the objective tests.

Method

Participants

An ad was placed in the school district's internal newsletter and a bulletin was mailed to all schools and offices requesting bilingual-certified volunteers to participate in a validation of the Spanish/English written test. Sixty-three individuals volunteered to participate. Forty-five of them completed two forms of the objective test, and the remaining 18 completed one form only. Volunteers included incumbents from 30 different job classifications, representing 11 nations of origin.

Predictor

Four equivalent forms of a multiple choice Spanish/English proficiency test were developed. Each form was divided into four sections: "Words in Context," a vocabulary test in which items consist of a sentence in Spanish or English containing an underlined word and for which testees must select the letter that begins the word in the other language; "Synonyms," an all-Spanish section in which testees select the word that means most nearly the same as the stimulus word; "Antonyms," an all-Spanish section in which the opposite of the stimulus word is selected; and "Translations," which consists of two passages, one in each language, that contain underlined phrases or sentences for which testees must select the correct translation. All items in the tests were developed using materials actually translated on the job by bilingual personnel of the district. Alpha coefficients for the whole tests and subtests of the Spanish/English written bilingual exam (obtained from another administration) are presented in Table 1.

Insert Table 1 about here

Those participants who completed two forms of the exam completed either Forms A and C or Forms B and D. Test administration was counterbalanced for order of presentation.

Criterion

Prior to completing the objective tests, each participant was asked to translate, longhand and without the benefit of dictionaries, a passage from written Spanish into written English. The stimulus passage was the same one contained in the Translations section of the form of the written test each would subsequently take.

Four raters were selected to evaluate the longhand translations completed by the volunteers. The raters were selected based on their demonstrated bilingual proficiencies in Spanish and English. One was an official translator of the district, and the other three were certificated personnel who had taught Spanish at the secondary school level or higher.

The raters evaluated the translations using a holistic scoring guide (Appendix A). The guide was similar to those described by White (1986) and utilized by Honey (1990) to score original creative writing samples. Numerical scores ranged from one to six and descriptive scores ranged from "Incompetent" (one) to "Superior" (six). Prior to scoring the actual translations produced by the study participants, the raters were trained using sample translations exemplifying each of the numerical anchors. The raters were then paired in teams of two to evaluate each translation. The first rater printed the translation's score on the back of the translation so as not to influence the score of the other rater. One rater team evaluated translations associated with Forms A and B, the other team evaluated translations taken from Forms C and D. In 45 cases, this meant that writing samples from individual participants were evaluated by all four raters.

Data Analysis

Reliability measures and validity coefficients were obtained using the Pearson r correlational coefficient.

Results

Table 2 displays the interrater reliability measures obtained using the holistic scoring procedure. As can be seen, these results indicate that the average interrater reliability obtained in the study was .69, and

that the two rater teams had interrater reliability measures of .71 and .67.

Insert Table 2 about here

A whole test reliability measure was also obtained for those individuals who completed two forms of the bilingual proficiency test. The correlation between test scores for these 45 individuals was .81.

Table 3 presents the results of the correlations between scores obtained in the rater evaluations of freehand translations and objective test scores. Table 3 indicates that objective test scores correlated with rater scores for Raters 1 and 2 at .69; objective test scores correlated with scores obtained from Raters 3 and 4 at .47. For those participants who completed two forms of the bilingual proficiency test (and thus also two translations), the average score obtained by the individual on the objective tests and the average score assigned by the four raters for the translations produced a correlational coefficient of .62.

Insert Table 3 about here

All correlations obtained in the data analysis were statistically significant at $p < .001$.

Conclusions

An initial concern with conducting a validation study using the procedure described above was that volunteer participants might not be representative of the range of abilities existing among bilingual employees; that is, individuals with relatively poorer bilingual skills

might be unlikely to participate and that therefore the range of scores observed in the study would be severely restricted. This did not occur. Figure 1 presents a scatterplot of average objective test score versus average holistic rating, and graphically demonstrates that a wide range of abilities was detected with both measures.

Insert Figure 1 about here

The relationship between objective test scores and holistic evaluations of translations was used to establish a passing point for the objective tests. The pass point was drawn at 75%, a level which would have caused four of the study participants to fail the exam. Two of these four were also the lowest scores obtained from the raters, averaging less than "unacceptable" and somewhat higher than "incompetent." The average ratings of all four participants were less than "inadequate."

The interrater reliabilities observed in the present study are somewhat lower than some reported in the literature for studies utilizing a holistic approach to the evaluation of creative writing. For example, White (1986) discusses interrater reliability measures obtained in a study of writing evaluation and its replication one year later. Utilizing a research design very similar to that employed in the present study, groups of four trained raters assigned scores on a six-point scale to essays written by college students. Each of the nearly 700 papers evaluated by the raters contained two essays composed by a single individual, and each essay was evaluated by two raters. The scores for both essays were combined to produce a single score for each student. White obtained interrater reliability coefficients of .85 and .78 for the initial study

and the replication, respectively. Cooper (1977) also cites numerous studies of holistic evaluation indicating interrater reliabilities ranging between .73 and .95.

Several explanations can be offered for the lower reliability observed in the present study. The first of these concerns rater training. The SMEs in the present study received a single exposure to anchor papers prior to the scoring of actual translations. In the studies reported by White (1986), raters received more extensive training; typically, at least one more set of sample anchor papers was scored before actual scoring began. Perhaps if our raters had had another set of samples for comparison following the discussion of the first set, they may have exhibited greater correspondence.

Another explanation concerns rater similarity, that is, the degree of similarity in training and background among the raters. As Cooper (1977) has noted, the more similar raters are in terms of background, the more likely it is that high reliability coefficients will be obtained for the evaluation of writing. It is possible that dissimilarity of raters is a problem exacerbated by the task of evaluating an individual's ability to translate from one language into another: raters must not only have a similar frame of reference for the first language, but also for the second. It may be that equally competent translations are judged better or worse because of regional preferences for one style over another, for example.

A last explanation concerns the methods used to compute the reliability estimates reported in the literature. As White observes:

> ". . . Unfortunately, statisticians have reached no agreement about ways to measure comparative reliability of readings, and there

> are many different ways of computing reading reliability. Reports of reliabilities in the .90 range occur from time to time (for example in the opening chapter of Cooper and Odell, 1977), but it is rarely clear how these numbers are derived. The statistical report published each year by the California State University English Equivalency Examination includes a statistical consultant's narrative of the methodology and shows just how the reliability statistics have been developed. The report on the 1981 test concludes that 'the actual reliability of the essay test therefore lies between the lower limit of .68 and the upper limit of .89' . . . In general, it appears that carefully controlled essay readings do yield reasonably reliable scores, although a close reading of the data reported by many such readings suggests that overstatement of reliability is common."

In sum, it appears that the interrater reliability measures obtained in the present study are consistent with the range of interrater reliabilities reported in the literature, albeit at the lower end of the range.

Valdes (1989) has summarized the evolution to date of testing for bilingual proficiency. She breaks this history into three parts: a "pre-scientific" stage, during which tests were concerned primarily with assessing the written grammar and translation competencies of those who had acquired bilingual ability through formal instruction; a "discrete points"

stage, during which tests were developed to measure discrete components of linguistic ability individually; and the "global" or "integrative" phase through which we are now proceeding, and in which test developers are attempting to assess the bilingual individual's ability to use many linguistic elements simultaneously. The strength of the relationship observed between objective test scores and raters' evaluations supports the conclusion that the holistic scoring of translations can serve as an adequate validation method for the objective test, and as such makes a contribution toward the expansion of global testing for bilingual proficiency.

REFERENCES

Cascio, W. (1986). Technical and mechanical job performance appraisal. In: Performance Assessment: Methods and Applications, R.A. Berk (Ed.), The Johns Hopkins University Press: Baltimore, MD.

Center for Continuing Study of the California Economy (1990) California Population Characteristics. Palo Alto, CA.

Cooper, C.R. (1977). Holistic evaluation of writing. In: Evaluating Writing: Describing, Measuring, Judging, C.R. Cooper and L. Odell (Eds.), National Council of Teachers of English: Urbana, IL.

Honey, R.A. (1990). The assessment of writing ability: The effectiveness of different models and formats. Paper presented at the Fourteenth Annual Conference of the International Personnel Management Association Assessment Council, San Diego, CA.

Lawton, J.R. and B.A. Showers (1980). Writing assessments for Merit System employment. Paper presented at the Fourth Annual Conference of the International Personnel Management Assessment Council, Boston, MA.

Quellmalz, E.S. (1986) Writing skills assessment. In: Performance Assessment: Methods and Applications, R.A. Berk (Ed.), The Johns Hopkins University Press: Baltimore, MD.

Valdes, G. (1989). Testing bilingual proficiency for specialized occupations. In: Test Policy and Test Performance: Education, Language, and Culture, B.R. Gifford, (Ed.), Kluwer Academic Publishers: Boston, MA.

White, E.M. (1986). Teaching and Assessing Writing, Jossey-Bass Publishers: San Francisco, CA, pp. 26-27, 160, 178-180.

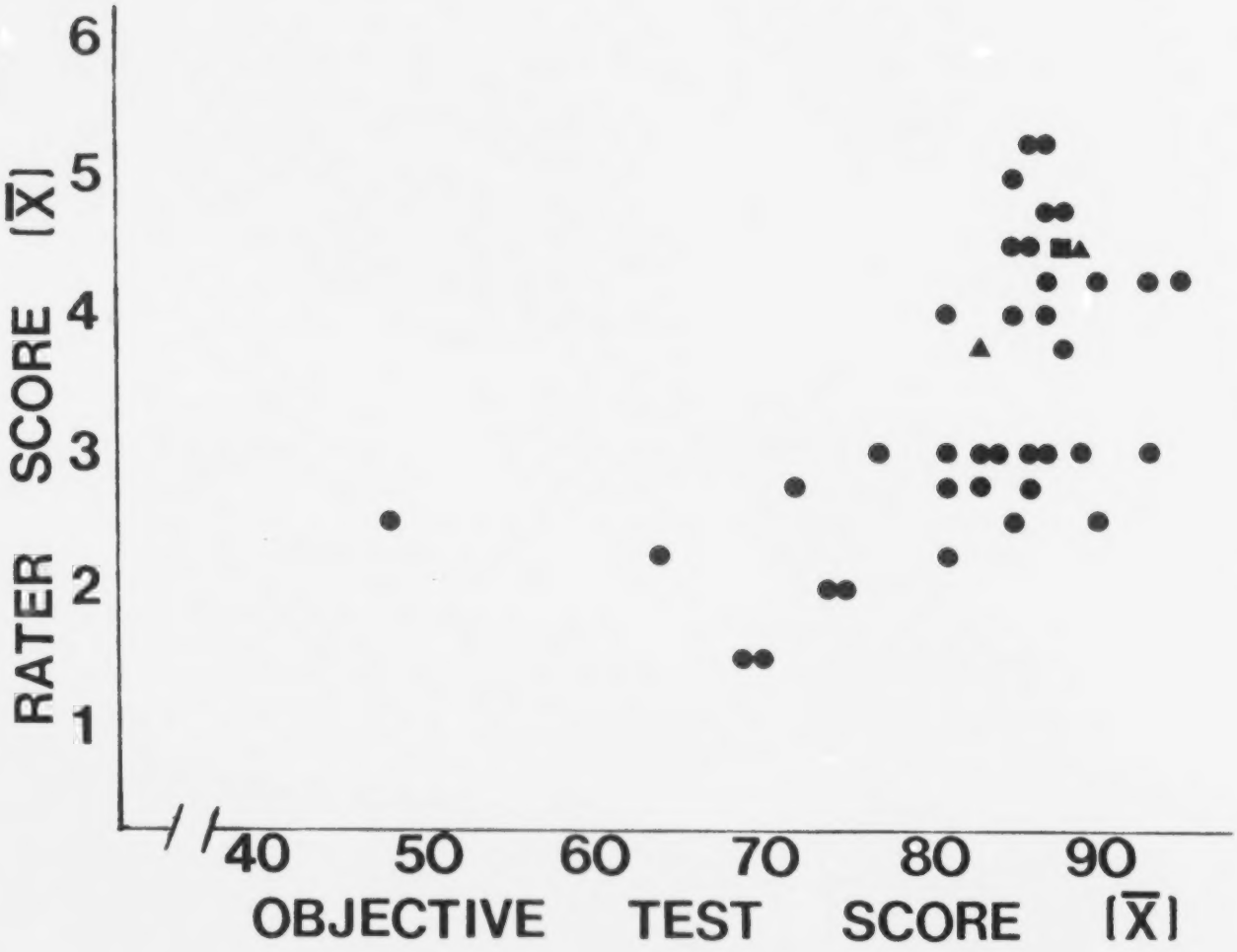

Figure 1. Scatterplot for objective test scores and translation ratings assigned by SMEs. ● = 1 data point; ▲ = 2 data points; ■ = 4 data points.

Table 1.

1991 Spanish/English Bilingual Written Test
Alpha Reliability Coefficients

	FORM A (N=1224)	FORM B (N=425)	FORM C (N=830)	FORM D (N=371)
WHOLE TEST (95 items)	.94	.84	.88	.89
WORDS IN CONTEXT (25 items)	.72	.66	.71	.65
SYNONYMS/ANTONYMS (40 items)	.91	.69	.84	.83
TRANSLATIONS (30 items)	.80	.72	.74	.80

Table 2.

Interrater Reliabilities

Rater 1 with Rater 2	$r = .71$	$p < .001$	$N = 55$
Rater 3 with Rater 4	$r = .67$	$p < .001$	$N = 53$

Table 3.

Correlations between Objective Scores and SME Ratings of Translations

Score with Raters 1 and 2	$r = .69$	$p < .001$	$N = 55$
Score with Raters 3 and 4	$r = .47$	$p < .001$	$N = 53$
Average Score with Average Rating	$r = .62$	$p < .001$	$N = 45$

APPENDIX A

SPANISH TRANSLATION
HOLISTIC SCORING GUIDE

The following scoring guide should be used to evaluate the translations of an English paragraph into Spanish.

6 **A Superior Translation:**
- fluent, well-organized, and complete
- translation is accurate and faithful to the original text, without embellishments and/or omissions
- level of discourse corresponds to that of the original text
- vocabulary demonstrates facility and fluency in both languages
- relatively free of errors in sentence structure, grammar, and mechanics

5 **A Proficient Translation:**
- fluent, well-organized and complete, despite very occasional technical writing faults
- translation is accurate, though minor embellishments or omissions may be present
- level of discourse corresponds to that of the original text
- though there may be minor lapses that could be rectified easily through the use of a dictionary, vocabulary generally indicates facility in both languages
- any errors in sentence structure, grammar, or mechanics are minor and can be readily corrected

4 **A Basically Competent Translation:**
- demonstrates basic competence, but there are problems in sentence structure, word selection, or mechanics
- essential information is conveyed, but some relevant information may be omitted, or embellishments convey information not contained in the original text
- level of discourse may be adequate but not of the level of the original text
- lapses in vocabulary could be corrected with a dictionary
- parts of the translation need to be rewritten to ensure proper interpretation

SPANISH TRANSLATION
HOLISTIC SCORING GUIDE

(continued)

3 **An Inadequate Translation:**
- problems in sentence structure, word selection and/or mechanics contribute to poor readability
- important information may have been omitted, or embellishments could cause a misinterpretation
- level of discourse is well below that of the original text
- lapses in vocabulary indicate difficulty in one language
- major revision or rewriting is required

2 **An Unacceptable Translation:**
- generally lacks coherence
- probably omits much necessary information
- numerous errors in sentence structure, grammar, mechanics, and/or word selection
- numerous lapses in vocabulary indicate very poor skills in one language
- requires extensive rewriting

1 **An Incompetent Translation:**
- demonstrates clear incompetence
- not suitable for revision
- requires complete rewriting